STAT 534 Lecture 11
Markov Randon Fields and Estimating their parameters

©Marina Meilă
mmp@stat.washington.edu

1 What is a Markov Random Field?

An arbitrary undirected graph can be seen as encoding a set of independencies. Let V be the set of nodes of G, each representing a random variable, and $\mathcal{E}$ the set of edges. Denote $n(A)\;=\;$ the **neighbors** of variable A.

Then the **local Markov property** is expressed as

$$\boxed{A\;\perp\;\text{everything else}\mid n(A)}$$

Now we will characterize the set of distributions that satisfy the Local Markov Property w.r.t. a graph G. For this, we need a new definition. A **clique** of a graph G is a set of nodes $C\subseteq V$ which are **fully connected** in G (i.e all possible edges between nodes in C appear in $\mathcal{E}$). A **maximal** clique is a clique which is not contained in any other clique of the graph.

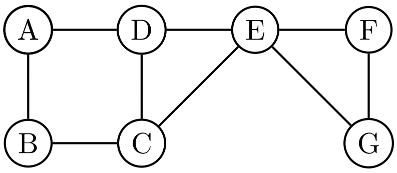
For example, in figure 1, all the nodes are cliques of size one (but not maximal), all the edges are cliques of size two, and the triangles CDE, EFG are cliques of size three. The maximal cliques are AB, BC, AD, CDE, EFG.

Theorem 1 *Let G be a graph and assume P can be factored in the following way*

$$P\;=\;\prod_{C\,\text{maximal clique}}\phi_C(x_C) \tag{1}$$

where ϕ_C is a non-negative function depending only on the variables in C. Then, P satisfies the Local Markov Property w.r.t. G

We will illustrate this theorem by an example shortly. The converse is a more powerful result, and is known as the Hammersley-Clifford theorem.



Examples: $F, G \perp A, B, C, D \mid E$
$A \perp C \mid B, D$
$A \perp C \mid B, D, E, F$

Figure 1: An undirected graph and some independencies encoded by it.

Theorem 2 (Hammersley-Clifford) *If $P > 0$ and P satisfies the Local Markov Property w.r.t G, then P can be written as a product of functions defined over the cliques of G as in (1)*[1]*.*

Exercise The theorem doesn't always hold if $P(x) = 0$ for some x. Can you construct such a counterexample? (Hint: give P lots of zeros.)

If a distribution P can be written in the form (1) for some graph G we say that P **factors according to graph** G.

Example Factorization for the undirected G in figure 1

$$P_{ABCDEFG} = \phi_{AB}(a,b)\phi_{AD}(a,d)\phi_{BC}(b,c)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g)$$

The functions ϕ are called **clique potentials**. They are required to be non-negative (positive if $P > 0$). Clique potentials are not uniquely defined. One can obtain equivalent factorizations by dividing/multiplying with functions of variables that are common between cliques. For instance, we can rewrite the above joint distribution as

$$\begin{aligned}
&P_{ABCDEFG} = \\
&= (2\phi_{AB}(a,b))(\phi_{AD}(a,d)/2)\phi_{BC}(b,c)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g) \\
&= (h(a)\phi_{AB}(a,b))(\phi_{AD}(a,d)/h(a))\phi_{BC}(b,c)\phi_{CDE}(c,d,e)\phi_{EFG}(e,f,g) \quad \text{for any } h(a) > 0 \\
&= \Phi_{AB}(a,b)\phi_{AD}(a,d)\phi_{BC}(b,c)(\phi'_{CDE}(c,d,e)h(c,d))\phi_{EFG}(e,f,g)
\end{aligned}$$

The last example shows why we only need to consider maximal cliques in the factorization of P. Because of the non-unicity of the ϕ's, the parameters of the clique potentials are hard to interpret. The potentials do not, in general, represent probability tables. However, there are some important special cases when the ϕ's

[1] The original Hammersley-Clifford theorem is stronger; it only assumes that P obeys the local Markov property according to G

have probabilistic interpretations – these will be the decomposable models we will study later. The Hidden Markov model you have already encountered is one of them.

1.1 The clique potentials are not marginals in Markov Random Fields - an example

The following simple example shows that *potential* $\neq$ *marginal* even if the potential is normalized.

Let $V = \{A, B, C\}$, $\mathcal{E} = \{AB, BC, CA\}$ and

$$\phi_{AB} \;=\; \phi_{BC} \;=\; \phi_{AC} \;=\; \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

Note that this is not exactly a Markov field, as the potentials are given on the edges, not on the maximal clique ABC. However, we shall use this example for simplicity (otherwise we'd need 4 or more nodes to prove our point). Namely, we will show that $P_{AB} \not\propto \phi_{AB}$.

$$\begin{aligned} P_{AB}(0,0) &\propto \phi_{AB}(0,0)\sum_C \phi_{AC}(0,c)\phi_{BC}(0,c) = \frac{1}{3}\frac{1}{3}\frac{1}{3} + \frac{1}{3}\frac{1}{6}\frac{1}{6} = \frac{5}{3^3 \cdot 4} \\ P_{AB}(0,1) &\propto \phi_{AB}(0,1)\sum_C \phi_{AC}(0,c)\phi_{BC}(1,c) = \frac{1}{6}\frac{1}{6}\frac{1}{3} + \frac{1}{6}\frac{1}{3}\frac{1}{6} = \frac{2}{3^3 \cdot 4} \end{aligned}$$

By symmetry, $P_{AB}(1,1) = P_{AB}(0,0)$ and $P_{AB}(0,1) = P_{AB}(1,0)$. Hence

$$Z \;=\; 2(P_{AB}(0,0) + P_{AB}(0,1)) \;=\; \frac{7}{54}$$

and

$$\frac{P_{AB}(0,0)}{P_{AB}(0,1)} \;=\; \frac{5}{2} \neq \frac{\phi_{AB}(0,0)}{\phi_{AB}(0,1)} \;=\; \frac{2}{1}$$

2 Parameter estimation. The (log)-likelihood

The estimation will be considered only in the ML framework. Let

$$P_V = \frac{1}{Z}\prod_C \phi_C(x_C) \tag{2}$$

be a joint distribution represented as a Markov Random Field (MRF), with $\{C\}$ the set of cliques and Z the normalization constant.

The parameter estimation problem calls for the estimation of the entries in all the potential tables ϕ_C. We assume that each entry is a different parameter, and denote it (abusively, perhaps) by $\phi_C(x_C)$.

Denote by $\mathcal{D}$ a data set of complete observations of the variables in V, sampled i.i.d. from an unknown distributions. We denote by $N_C(x_C)$ the number of times configuration x_C over the variables in C appears in the data. We shall see that, just as in the case of the BN, the set of N_C counts are the sufficient statistics for the parameters.

We have that $\sum_{x_C \in \Omega_C} N_C(x_C) \ = \ N$.

Example Let $V = \{A, B, C, D\}$ with all variables taking values in $\{0, 1\}$ and

$$P_V \ = \ \frac{1}{Z}\phi_{AB}(a,b)\phi_{BC}(b,c)\phi_{CD}(c,d)\phi_{DA}(d,a) \tag{3}$$

with

$$Z \ = \ \sum_{a \in \Omega_A} \sum_{b \in \Omega_B} \sum_{c \in \Omega_C} \sum_{d \in \Omega_D} \phi_{AB}(a,b)\phi_{BC}(b,c)\phi_{CD}(c,d)\phi_{DA}(d,a) \tag{4}$$

(Thus, Z is a sum over 16 terms.) Let the data set contain $N = 5$ samples.

A	B	C	D
1	1	1	0
1	0	0	1
0	1	1	0
0	0	1	0
1	0	1	1

The log-likelihood is the logarithm of the probability of the data $\mathcal{D}$ under the model P_V, i.e

$$l(\text{parameters}; \mathcal{D}) \ = \ \sum_C \sum_{x_C \in \Omega_C} N_C(x_C) \ln \phi_C(x_C) - N \ln Z \tag{5}$$

The first observation we make is that the data enter the likelihood only via the *sufficient statistics* $N_C(x_C)$. Then, we will find it convenient to normalize both sides of the above equation by the smaple size N. The ratio $\frac{N_c(x_C)}{N} = \hat{P}_C(x_C)$ represents a probability, namely the *empirical distribution* of the variable(s) in clique C, or the *sample marginal* of C w.r.t. the empirical distribution represented by the sample $\mathcal{D}$. After the normalization we obtain

$$\frac{1}{N}l = \sum_C \sum_{x_C \in \Omega_C} \hat{P}_C(x_C) \ln \phi_C(x_C) - \ln Z \tag{6}$$

The normalization constant Z is a function of all parameters.

For the **example** above, we have

$$N_{AB} = \begin{array}{r|cc} B: & 0 & 1 \\ \hline A=0 & 1 & 2 \\ 1 & 1 & 1 \end{array} \qquad N_{BC} = \begin{array}{r|cc} C: & 0 & 1 \\ \hline B=0 & 1 & 2 \\ 1 & 0 & 2 \end{array} \qquad \ldots \tag{7}$$

The log-likelihood of this data is

$$\begin{aligned} \frac{1}{N}l &= \left(\frac{1}{5}\ln \phi_{AB}(0,0) + \frac{2}{5}\ln \phi_{AB}(0,1) + \frac{1}{5}\ln \phi_{AB}(1,0) + \frac{1}{5}\ln \phi_{AB}(1,1)\right) & (8) \\ &+ \left(\frac{1}{5}\ln \phi_{BC}(0,0) + \frac{2}{5}\ln \phi_{BC}(0,1) + \frac{2}{5}\ln \phi_{BC}(1,1)\right) + \ldots - \ln Z & (9) \end{aligned}$$

3 Maximizing the likelihood by gradient ascent

To find the maximum value for the parameters, we use the iterative procedure called **gradient ascent**.

```
GradientAscent
 Input sufficient statistics $N_C(x_C)$ (or sample marginals $\hat{P}_C(x_C)$) for all $C$ and all $x_C$
 Initialize $\phi_C$ with arbitrary values
     Repeat
         1. $\phi_C(x_C) \leftarrow \phi_C(x_C) + \eta \frac{\partial l/N}{\partial \phi_C(x_C)}$ for all $x_C$
     until "convergence"
```

In other words, GradientAscent iteratively corrects the current parameter estimates with a correction that will increase the log-likelihood l. The parameter $\eta > 0$ is a *step size* that is sometimes fixed and sometimes estimated at each step (as we shall see in STAT 538).

GradientAscent is a generic optimization algorithm. To use it we need to calculate the expression of the gradient $\frac{\partial l/N}{\partial \phi_C(x_C)}$. We do so now.

$$\frac{\partial l/N}{\partial \phi_{C_0}(x^*_{C_0})} = \frac{\partial}{\partial \phi_{C_0}(x^*_{C_0})} \sum_C \sum_{x_C \in \Omega_C} \frac{N_C(x_C)}{N} \ln \phi_C(x_C) - \frac{\partial}{\partial \phi_{C_0}(x_{C_0^*})} \ln Z \tag{10}$$

$$= \underbrace{\frac{N_{C_0}(x^*_{C_0})}{N}}_{\hat{P}_{C_0}(x^*_{C_0})} \frac{1}{\phi_{C_0}(x^*_{C_0})} - \frac{\frac{\partial Z}{\partial \phi_{C_0}(x^*_{C_0})}}{Z} \tag{11}$$

$$\frac{\partial Z}{\partial \phi_{C_0}(x_{C_0^*})} = \sum_{x_V} \frac{\partial}{\partial \phi_{C_0}(x^*_{C_0})} \left[\prod_{C' \neq C_0} \phi_{C'}(x_{C'}) \phi_{C_0}(x^*_{C_0}) \right] \tag{12}$$

$$= \sum_{x_{V \setminus C_0}} \prod_{C' \neq C_0} \phi_{C'}(x_{C' \setminus C_0}, x^*_{C' \cap C_0}) \tag{13}$$

$$= \frac{Z}{\phi_{C_0}(x_{C_0^*})} \sum_{x_{V \setminus C_0}} \prod_{C' \neq C_0} \phi_{C'}(x_{C' \setminus C_0}, x^*_{C' \cap C_0}) \frac{\phi_{C_0}(x_{C_0^*})}{Z} \tag{14}$$

$$= \frac{Z}{\phi_{C_0}(x_{C_0^*})} \sum_{x_{V \setminus C_0}} P_V(x_{V \setminus C_0}, x^*_{C_0}) \tag{15}$$

$$= Z \frac{P_{C_0}(x^*_{C_0})}{\phi_{C_0}(x^*_{C_0})} \tag{16}$$

$$\frac{\partial l/N}{\partial \phi_{C_0}(x^*_{C_0})} = \frac{\hat{P}_{C_0}(x^*_{C_0}) - P_{C_0}(x^*_{C_0})}{\phi_{C_0}(x^*_{C_0})} \tag{17}$$

Thus, the gradient w.r.t a parameter $\phi_{C_0}(x^*_{C_0})$ depends on: (1) the current value of the parameter $\phi_{C_0}(x^*_{C_0})$, (2) the *empirical marginal* probability $\hat{P}$ of the configuration $x^*_{C_0}$, and (3) the marginal of the respective configuration as computed by the model P_V, i.e. $P_{C_0}(x_{C_0^*})$.

The first two quantities are readily available. The marginal $P_{C_0}(x_{C_0^*})$, however, must be computed. As we know, computing marginals is *inference*, thus we must be able to perform inference in the MRF in order to estimate the parameters.

For inference we use MCMC, which produces approximate marginals.

4 Iterative Proportional Fitting (IPF)

IPF is an alternative to gradient ascent, that does not require setting the step size.

The idea is that, at the optimum, the gradient will be zero. Hence we will have

$$\frac{\hat{P}_C(x_C)}{\phi_C(x_C)} \;=\; \frac{P_C(x_C)}{\phi_C(x_C)} \tag{18}$$

or

$$\phi_C(x_C) \;=\; \phi_C(x_C)\frac{\hat{P}_C(x_C)}{P_C(x_C)} \tag{19}$$

for all cliques C and configurations x_C. The IPF algorithm tries to reach this equilibrium point by multiplicative updates to the parameter $\phi_C(x_C)$ (while gradient ascent performs additive updates).

```
IPF **Algorithm**
Repeat
    for every clique $C \in \mathcal{C}$
        $\phi_C(x_C) \;\leftarrow\; \phi_C(x_C)\frac{\hat{P}_C(x_C)}{P_C(x_C)}$        (20)
until convergence
```

Proposition 1 The IPF algorithm preserves the value of the normalization constant Z.

Proof Assume that at step t the parameters of clique C are updated while the other cliques' parameters stay the same.

$$\begin{aligned}
P_C^{(t+1)}(x_C) &= \sum_{x_{V\setminus C}} P_V^{(t+1)}(x_V) && (21)\\
&= \sum_{x_{V\setminus C}} \prod_{C'} \phi_{C'}^{(t+1)}(x_{C'})/Z^{(t+1)} && (22)\\
&= \sum_{x_{V\setminus C}} \phi_C(x_C)^{(t+1)} \prod_{C'\neq C} \phi_{C'}^{(t)}(x_{C'})/Z^{(t+1)} && (23)\\
&= \sum_{x_{V\setminus C}} \phi_C^{(t)}(x_C)\frac{\hat{P}_C(x_C)}{P_C^{(t)}(x_C)} \prod_{C'\neq C} \phi_{C'}^{(t)}(x_{C'})/Z^{(t+1)} && (24)
\end{aligned}$$

$$= \frac{\hat{P}_C(x_C)}{Z^{(t+1)}P_C(x_C)} \sum_{x_{V\setminus C}} \underbrace{\phi_C(x_C)^{(t)} \prod_{C' \neq C} \phi_{C'}^{(t)}(x_{C'})}_{P_V^{(t)}(x)Z^{(t)}} \tag{25}$$

$$= \frac{Z^{(t)}\hat{P}_C^{(t)}(x_C)}{Z^{(t+1)}P_C(x_C)} \sum_{x_{V\setminus C}} P_V^{(t)}(x) \tag{26}$$

$$= \frac{Z^{(t)}\hat{P}_C(x_C)}{Z^{(t+1)}P_C^{(t)}(x_C)} P_C^{(t)}(x_C) \tag{27}$$

$$= \frac{Z^{(t)}}{Z^{(t+1)}}\hat{P}_C(x_C) \tag{28}$$

Summing now both sides over $x_C \in \Omega_C$ and noting that $\sum_{\Omega_C} \hat{P}_C = \sum_{\Omega_C} P_C = 1$ we obtain $\frac{Z^{(t)}}{Z^{(t+1)}} = 1$, which completes the proof.

From $Z^{(t)} = Z^{(t+1)}$ and (30) we can immediately derive:

Proposition 2 After updating the parameters ϕ_C we have that $P_C = \hat{P}_C$.

Hence, the IPF updates can be thought of as updating each clique iteratively in a way that makes its marginal equal to the data marginal. Next, we will give an interpretation for IPF as gradient ascent.

Let us consider the gradient (11) with the second term expressed by equation (15).

$$\frac{\partial l^{(t)}/N}{\partial \phi_{C_0}(x^*_{C_0})} = \frac{\hat{P}_{C_0}(x^*_{C_0})}{\phi_{C_0}(x^*_{C_0})} - \frac{P_{C_0}^{(t)}(x^*_{C_0})}{\phi_{C_0}^{(t)}(x^*_{C_0})} \tag{29}$$

Note that the first occurence of $\phi_{C_0}(x^*_{C_0})$ is as a *function argument*, while its second occurence is as a *parameter value*. We evaluate this gradient now at the *next* parameter value, $\phi_{C_0}^{(t+1)}(x^*_{C_0})$.

$$\left.\frac{\partial l^{(t)}/N}{\partial \phi_{C_0}(x^*_{C_0})}\right|_{\phi_{C_0}^{(t+1)}(x^*_{C_0})} = \frac{\hat{P}_{C_0}(x^*_{C_0})}{\phi_{C_0}^{(t+1)}(x^*_{C_0})} - \frac{P_{C_0}^{(t)}(x^*_{C_0})}{\phi_{C_0}^{(t)}(x^*_{C_0})} \tag{30}$$

If we set the condition that the gradient in this point is zero, we obtain equation (22). Setting this condition implies that we move in parameter space in the direction of the gradient until we (approximately) find a point where the gradient is zero, i.e. the point of the maximum increase along the gradient direction.